TRAM: Fine-Tuning-Free Test-Time Adaptation with Few Bonafide Samples for Generalized Face Anti-Spoofing

Qirui Wang[1,2,*] and Si-Qi Liu[1,3,*,✉]

[1] National Health and Medical Big Data Research Institute (Shenzhen), China
[2] School of Science and Engineering, The Second Affiliated Hospital, The Chinese University of Hong Kong, Shenzhen, China
[3] Shenzhen Research Institute of Big Data, China
qiruiwang@link.cuhk.edu.cn, siqiliu@sribd.cn

Abstract. Generalized face anti-spoofing (FAS) has attracted increasing attention due to the need for robustness in unseen scenarios. Domain adaptation methods can improve performance by leveraging target-domain information. However, in FAS, collecting spoof samples is significantly more expensive than acquiring a few bonafide samples, making it impractical to gather diverse attack types and finetune models for each deployment scenario. To address this challenge, we propose **T**ext-guided **R**el**A**tionship **M**odeling (TRAM), a finetuning-free test-time domain adaptation approach for generalized FAS using only a few bonafide samples. Since samples within a target domain share common domain factors, we treat genuine sample as the anchor and model its relationship with incoming inputs to mitigate domain bias. To learn discriminative spoof cues from these relationships, we introduce a Relative Image-Pair-Text Contrastive Learning strategy that leverages fine-grained FAS relation prompts. Furthermore, to compensate for the lack of detailed spoof annotations, we design a Fine-grained Relation Prompt Learning strategy that mines low-level photometric discrepancies through precomputed measurements to generate robust prompts. Extensive experiments on eight FAS datasets with typical variations demonstrate that TRAM achieves consistent improvements over state-of-the-art methods while using only a few bonafide samples from the target domain (as few as five). Comprehensive ablation studies further validate the effectiveness of each component.

Keywords: Face Anti-spoofing · Face presentation attack detection · Face liveness detection

1 Introduction

Face recognition systems are increasingly deployed across a wide range of applications, raising growing security concerns due to the prevalence of spoofing

* Equal contribution, ✉ Corresponding author

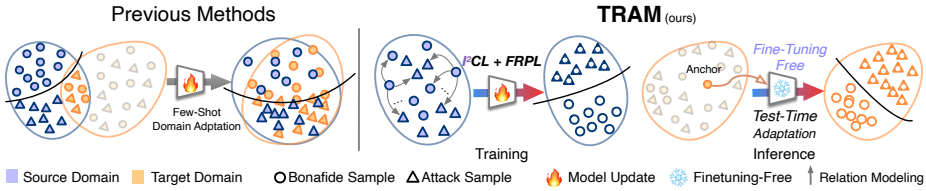

Fig. 1: Comparison between conventional domain adaptation (DA) methods for FAS and the proposed TRAM. Previous methods require both bonafide and attack samples from the target domain and perform model updating for distribution alignment. In contrast, TRAM enables finetuning-free test-time adaptation with only a few target-domain bonafide samples by modeling the relationship between an anchor sample and incoming inputs.

attacks. Such attacks can often be performed at very low cost by obtaining a user's facial image from social media and presenting it to the system via printed photos or replayed videos. To mitigate this threat, numerous face anti-spoofing (FAS) methods have been proposed, ranging from early hand-crafted feature approaches [1,21,43] to recent deep learning-based representation learning techniques [60,64,81]. While these methods achieve promising performance in intra-dataset evaluations, they often struggle to generalize to unseen domains. This limitation largely arises because FAS models rely on subtle cues such as texture patterns and color tone, which are highly sensitive to variations in imaging devices, illumination, and capture conditions, leading to significant domain shifts [33].

To address this challenge, domain generalization (DG) approaches aim to learn domain-invariant representations from multiple source domains [31,54,81]. Typical strategies include adversarial learning [22, 49] and meta-learning [5, 50], which attempt to improve robustness to unseen environments by reducing domain-specific biases during training. Recently, vision-language models (VLMs) such as CLIP have shown promising generalization ability in FAS [52] by aligning visual and textual representations in a shared embedding space. Subsequent works further improve robustness through prompt learning [28, 33] and fine-grained semantic guidance [33]. Alternatively, domain adaptation (DA) methods relax the DG assumption by allowing access to target-domain samples during adaptation. By explicitly reducing the distribution gap between source and target domains, DA methods often achieve stronger performance in practical deployment scenarios. Representative approaches include unsupervised DA that aligns feature distributions via discrepancy minimization [25] or adversarial learning [62], semi-supervised DA that leverages partially labeled target data [23, 47], and few-shot test-time adaptation that updates models using limited target samples [18,36,38,42,45].

However, these approaches usually require both bonafide and attack samples from the target domain, which increases deployment costs and may be infeasible in real-world scenarios. Collecting diverse attack samples across domains

is particularly challenging, as generating print or replay media or fabricating 3D masks requires manual effort and controlled environments. Moreover, adapting models with target data typically involves additional training or finetuning, which introduces extra computational overhead and limits scalability in online or large-scale systems.

This raises an important question: *Can we adapt FAS models to a new domain using only target domain bonafide samples?* In this work, we propose **T**ext-guided **R**el**A**tionship **M**odeling (TRAM), a CLIP-based finetuning-free test-time adaptation framework for FAS that adapts to new domains using only a few target-domain bonafide samples at inference without requiring attack samples. Our key insight is that genuine faces captured in the target environment share the same domain-specific factors (e.g., camera characteristics, lighting conditions, and image quality) with other incoming samples. By treating a genuine face as an anchor, TRAM compares each input sample against the anchor to suppress shared domain bias and highlight discriminative spoof cues. To achieve this, different from CLIP aligns image to its own text description, we design a relational text prompt that encourages the model to investigate the difference between the anchor and each input sample. For genuine face, the text prompt is simply No Difference. For attack sample, we explicitly describe the fine-grained spoofing artifacts (e.g., contrast, color saturation) as the relation prompts to better guide the model toward detecting FAS cues. We compose an anchor-relative FAS encoder to distill these relational features based on cross-correlation between the anchor and the input visual feature. A Relative Image-Pair-Text Contrastive Learning (I^2CL) framework is introduced to model anchor-relative relationships for spoof detection. Since fine-grained artifact descriptions may be missing or heterogeneous across different attack instruments, I^2CL is designed to handle incomplete artifact prompts and is regularized with consensus attention guidance. Furthermore, due to the lack of detailed spoof annotations, we propose a Fine-grained Relation Prompt Learning (FRPL) strategy that leverages pseudo artifact labels to mine low-level photometric discrepancies and learns robust relational prompts with conditioned attribute-balanced supervision to further guide I^2CL.

In sum, the main contributions of this work are:

- We propose TRAM, a novel CLIP-based finetuning-free test-time adaptation framework for face anti-spoofing that adapts to unseen target domains using only a few bonafide samples, eliminating the need for attack samples or additional training.
- TRAM introduces an image-pair relational modeling paradigm for FAS, consisting of 1) a Relative Image-Pair-Text Contrastive Learning (I^2CL) framework with consensus attention regularization to capture relational spoof cues between an anchor bonafide face and incoming inputs, and 2) a Fine-grained Relation Prompt Learning (FRPL) module that mines photometric discrepancies to generate detailed relational prompts with conditional attribute-balanced control.

– Extensive experiments on eight public FAS datasets demonstrate that TRAM achieves consistent improvements over state-of-the-art methods in cross-domain generalization using only a few bonafide samples from the target domain and without finetuning.

2 Related Work

2.1 Face Anti-Spoofing

Early FAS methods based on handcrafted features [3, 44, 55] evolved into deep learning-based frameworks [35, 37, 70] with the development of convolutional neural networks. Auxiliary tasks like depth maps [49], rPPG signals [32], and 3D structures [15] modeling are introduced to further enhance the discriminability. Despite the success in intra-dataset settings, these methods struggle to generalize to unseen domains due to domain shifts. While domain generalization techniques seek to learn domain-invariant representation [5,31,50,54,81], the effectiveness is constrained by the lack of target domain information, motivating the exploration of domain adaptation for FAS.

Domain Adaptation for FAS. Domain adaptation (DA) aims to align the source and target feature distributions to minimize the impacts of domain gaps. Early DA methods in FAS require access to both source and target data for joint training [16, 25, 61, 62]. This could be impractical in real-world scenarios due to privacy constraints and deployment costs. Source-free DA [39, 42] and later continue-learning based methods [4, 18] aim to solve this problem. Test-time DA (TTA) further lowers the costs since they can achieve distribution alignment without label of target domain samples [20, 38]. TTA can be conducted in a finetuning-free way by summarizing domain statistical information from inference batches [80], or translating [82]/projecting [26] the target domain images into source domain. However, these methods either need attack samples from target domain or restrict to the assumptions that domain shift in FAS can be regarded as style transfer. In this work. the proposed TRAM can achieve finetuning-free TTA without requiring any attack samples from the target domain. TRAM explicitly models relational FAS features in a fully learnable manner, by using only genuine face from the target domain as anchor and comparing it with incoming input samples.

Vision-Language-model-based FAS. CLIP has shown remarkable generalization by aligning image–text pairs in a joint embedding space. While originally designed for high-level vision tasks like classification and retrieval, its effectiveness in low-level tasks like denoising [8], defocus deblurring [68], and depth estimation [75] is also validated. FLIP was the first that introduces CLIP for FAS, achieving outstanding performance with text guidance: This is a `{real/spoof}`face [52]. Later works focus on improving generalizability through prompt engineering, including build learnable prompts [28,33] and constructing fine-grained descriptions with the help of LLMs [11,65,69,83]. Recent studies explored MLLM-based

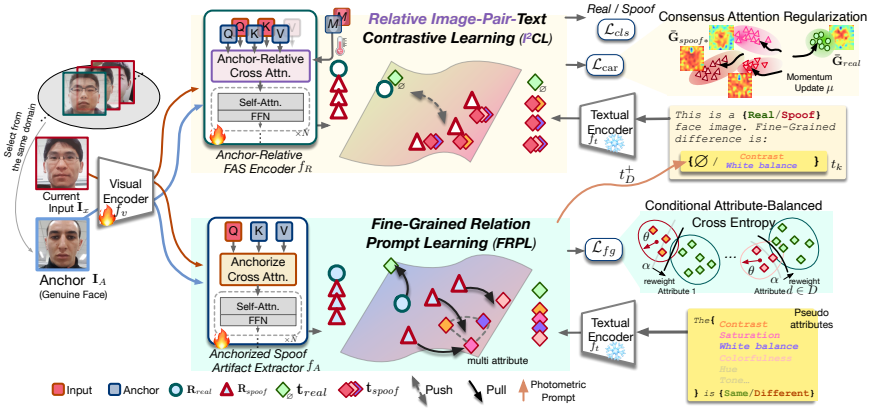

Fig. 2: Overview of the proposed method. TRAM takes an image pair as input, where one image serves as an anchor bonafide face. In Relative Image-Pair-Text Contrastive Learning (I^2CL), an anchor-relative FAS encoder models the relationship between the anchor and the input image under textual guidance $\boldsymbol{t}_k$, with consensus attention regularization. Fine-grained Relation Prompt Learning (FRPL) mines pseudo-artifact attributes to generate fine-grained prompts $\boldsymbol{t}_D^+$ that provide additional guidance for I^2CL.

FAS for improved interpretability [40,63]. In contrast to these DG settings without target-domain information, TRAM reconfigures CLIP for domain adaptation in FAS and achieve finetuning free test time adaptation with a few genuine samples.

2.2 Image Difference Captioning

Image Difference Captioning refers to generating descriptions that highlight the differences between two similar images. It mainly applies in the field of Remote Sensing [6,29] and viewpoint change tasks [46,56,57]. However, these tasks focus on explicit instance-wise changes and require accurate descriptions of differences for model training, which makes it difficult to integrate them to FAS. Furthermore, VisDiff [10] utilizes the LLM to generate text for image pairs separately and then determines the image differences based on the textual variances. However, the LLM lacks the core knowledge of FAS [51], which limits the textual differences generated to descriptions of instances, making it challenging to use for FAS classification. Typical cases are demonstrated in supplementary material.

3 Methodology

Overview. Figure 2 shows the overview of the proposed TRAM that takes an image pair as input—one is the input sample and the other is the anchor real face of the domain. TRAM conducts FAS by modeling the relationship between them via Image-Pair-Text contrastive learning. The network extracts relational

features with the proposed anchor-relative FAS encoder, which is optimized with the Relative Image-Pair-Text Contrastive Learning (I^2CL). These relational features are then matched with detailed relation prompt text features to conduct the final FAS classification. To provide such fine-grained relation prompts, the Fine-Grained Relation Prompt Learning (FRPL) module with an anchorized spoof artifact extractor is further introduced to mine FAS-relevant photometric discrepancy information and generate detailed spoofing-artifact prompts for I^2CL. During training, the anchor image is randomly sampled from bonafide samples in the source domain. During inference, a few bonafide samples from the target domain are used as anchor candidates to cover potential domain variations, and each incoming sample is paired with one of the anchor candidates for prediction.

3.1 Relative Image-Pair-Text Contrastive Learning

For the pair of input image $\boldsymbol{I}_x$ and the anchor $\boldsymbol{I}_A$ from the same domain with identical domain-specific information, we could get rid of the FAS-irrelevant domain-specific information and conduct generalized FAS by only looking at their differences. Therefore, we propose a Relative Image-Pair-Text Contrastive Learning (I^2CL) framework to learn generalized FAS information by investigating their relational features with text guidance. Specifically, an anchor-relative FAS encoder f_R is constructed on top of the CLIP visual encoder to model the relationship of their visual feature $\boldsymbol{x}$ and $\boldsymbol{A}$:

$$\boldsymbol{R} = f_R(\boldsymbol{A}, \boldsymbol{x}) \tag{1}$$

Different from the previous CLIP-based methods that take one image as the input, here the text guidance of $\boldsymbol{R}$ is formed with relationship description, i.e.: "`This is a {real / spoof} face. The face is {the same as / different from} the anchor.`" Then the final FAS classification results $\hat{y}$ is obtained by selecting the class prompt $\boldsymbol{t}_r$ and $\boldsymbol{t}_s$ with the highest response for $\boldsymbol{R}$:

$$\hat{y} = \underset{k\in\{\mathrm{r,s}\}}{\operatorname{argmax}} \langle \boldsymbol{R}, f_t(\boldsymbol{t}_k) \rangle \tag{2}$$

where f_t is the text encoder and $\langle , \rangle$ denotes cosine similarity. We follow [52] and set CLIP loss $\mathcal{L}_{cls} = CE(y, p(\hat{y}|\boldsymbol{R}))$, where y denotes the label of $\boldsymbol{I}_x$.

Anchor-Relative Cross Attention. Given two input features $\boldsymbol{x}$ and $\boldsymbol{A}$, we employ cross attention mechanism to model their relationship. f_R is constructed in the ViT manner where the first layer we design the Anchor-Relative Cross Attention (RCA). By applying transformer linear mapping [59], we obtain $\boldsymbol{Q}_x$, $\boldsymbol{K}_x$,$\boldsymbol{V}_x$ from $\boldsymbol{x}$ and $\boldsymbol{Q}_A$, $\boldsymbol{K}_A$, $\boldsymbol{V}_A$ from $\boldsymbol{A}$. To suppress the shared non-FAS characteristics between $\boldsymbol{x}$ and $\boldsymbol{A}$, we use the difference of their $\boldsymbol{Q}$ and $\boldsymbol{K}$ to remove common facial attributes and background factors, allowing the model to focus on spoof-related cues in relationship modeling. RCA is formulated as:

$$\begin{aligned} &\mathrm{RCA}(\boldsymbol{Q}_A, \boldsymbol{Q}_x, \boldsymbol{K}_A, \boldsymbol{K}_x, \boldsymbol{V}_A) = \\ &\qquad \mathrm{softmax}\Big(\big(\boldsymbol{Q}_A - \boldsymbol{Q}_x\big)\big(\boldsymbol{K}_A - \boldsymbol{K}_x\big)^\top \odot \mathcal{T}\big(\boldsymbol{M}_x^T \boldsymbol{M}_A\big) \Big) \boldsymbol{V}_A \end{aligned} \tag{3}$$

Note that we modulate the attention temperature with facial prior masks $\boldsymbol{M}_x$, $\boldsymbol{M}_A$ to emphasize facial areas with more generalized FAS cues. $\mathcal{T}(\cdot)$ is a modulation function. The output of RCA is followed by Self-attention and FFN layers to form the entire f_R.

Consensus Attention Regularization. Empirically, CLIP tends to exhibit sparse attention behaviors [79]. For FAS such sparsity may cause overfitting on domain-specific spoofing cues, e.g., the handhold prints attack or 3D mask with eye region artifacts, which could hinder the learning of generalized relational features. To mitigate this issue, we propose Consensus Attention Regularization(CAR) to encourage samples from real faces and each attack type to form their own consistent relational-attention distribution.

Specifically, for each category c of real class and different attack types, we construct a consensus attention map $\bar{\boldsymbol{G}}_c$. Given an input pair $(\boldsymbol{A_i}, \boldsymbol{x_i})$, we obtain its Grad-CAM [48] map $\boldsymbol{G}_i^c$ from f_R and update the corresponding $\bar{\boldsymbol{G}}_c$ with an exponential moving average:

$$\bar{\boldsymbol{G}}_c^{(n+1)} = \mu \bar{\boldsymbol{G}}_c^{(n)} + (1-\mu)\boldsymbol{G}_i^c, \tag{4}$$

where μ controls the momentum of the update and n denotes the update step index. Through iterative updates, $\bar{\boldsymbol{G}}_c$ gradually reflects the consensus relational attention pattern within each category. Once the $\bar{\boldsymbol{G}}_c$ are stabilized , they serve as targets that tune the model's relational attention by minimizing the divergence between each sample's attention map and its corresponding $\bar{\mathbf{G}}_c$:

$$\mathcal{L}_{\text{car}} = \text{JS}(\boldsymbol{G}_i^c, \bar{\boldsymbol{G}}_c^{(n)}), \tag{5}$$

where $\text{JS}(\cdot)$ denotes the Jensen–Shannon divergence.

3.2 Fine-grained Relation Prompt Learning

To better describe the relationship between the $\boldsymbol{x}$ and $\boldsymbol{A}$ and make the text guidance precise and specific, we propose to incorporate the fine-grained spoofing artifacts into the text prompts for I^2CL. As validated in [25], the spoofing samples tend to have different photometric properties compared with the genuine one due to the quality defects of spoofing medium, such as the contrast, saturation, colorfulness and white balance. These discriminative photometric properties can be adopted as the fine-grained spoofing artifacts attributes to form the detailed relationship description in I^2CL. Since these detailed attributes are not available for existing datasets, intuitively, we may extract photometric features and use numerical statistics with a certain threshold to obtain them. However, since photometric differences can also exist in genuine sample pairs due to variations of lighting conditions or skin color, such hand-crafted statistics may introduce noisy labels issue and lead to catastrophic performance degradation. To avoid this problem, we design the Fine-Grained Relation Prompt Learning (FRPL) on top of the pre-computed photometric attributes to learn robust fine-grained

prompts. FRPL employs the anchorized spoof artifact extractor to obtain artifact representations and selects the relevant fine-grained prompts by measuring their similarity in the embedding space.

Anchorize Cross Attention. As shown in Fig. 2, the anchorized spoof artifact extractor f_A is constructed to mine low-level FAS-related photometric discrepancy information $\boldsymbol{R}_A$:

$$\boldsymbol{R}_A = f_A(\boldsymbol{A}, \boldsymbol{x}) \tag{6}$$

Inspired by [57], we perform cross-attention on $\boldsymbol{Q_x}$ and $\boldsymbol{K_A}$ to find the spoof artifact feature and form the Anchorize Cross Attention (ACA) as:

$$ACA(\boldsymbol{Q}_x, \boldsymbol{K}_A, \boldsymbol{V}_A) = softmax(\boldsymbol{Q}_x \boldsymbol{K}_A^T \odot \mathcal{T}(\boldsymbol{M}_x^T \boldsymbol{M}_A))\boldsymbol{V}_A \tag{7}$$

where d_{K_A} is the dimension of $\boldsymbol{K_A}$. Similarly, f_A follows ViT style design which attaches ACA with several self-attention and FFN layers to obtain the output.

Attribute-balanced Multi-label Photometric Prompt Learning. Since spoof artifacts may exist in multiple photometric attributes for spoofing samples, we compose a CLIP-style multi-label classification [53] task to learn the prompt, i.e., composing two classes(same/different) for each photometric attribute and computing the similarities with $\boldsymbol{R}_A$. We compute photometric properties from the input image pair on the fly and analyze the difference to generate pseudo attribute label. The ratio of photometric property values of $\boldsymbol{I}_x$ and $\boldsymbol{I}_A$ is measured. If ratio r of property d is below a certain threshold θ and $\boldsymbol{I}_x$ is a spoof sample, we mark d as `{different}` and vice versa. Detailed metrics are in the supplementary material.

Since photometric attributes can also be affected by other factors, we set θ with a relatively low value to reduce the noise of pseudo labels. But this could lead to data unbalanced issues where the number of `{different}` can be significantly lower than that of `{same}`. To solve it, we compose a Conditional Attribute-Balanced Cross Entropy $\mathcal{L}_{fg}$ that can control pseudo-attribute-label noise and mitigate class imbalance effects simultaneously:

$$\begin{aligned}\mathcal{L}_{fg} = -\frac{1}{D}\sum_{d=1}^{D} \mathbf{1}_{\{\boldsymbol{r}_d<\theta \wedge y=1\}} \alpha \log(p(\hat{y}_d^+|\boldsymbol{R}_A)) + \\ \mathbf{1}_{\{\boldsymbol{r}_d\geq\theta \vee y=0\}}(1-\alpha)\log(1-p(\hat{y}_d^-|\boldsymbol{R}_A))\end{aligned} \tag{8}$$

where D is the number of photometric attribute classes, $\boldsymbol{r}_d$ denotes the ratio of d-th attribute, and $\hat{y}^+$ and $\hat{y}^-$ represent the prediction of `different` and `same`. α with the condition controls the class weight. For each input pair of a genuine sample and a sample of class y (real/fake), we compute $\mathcal{L}_{fg}$ so that the encoder $f_A(\cdot)$ can generate robust photometric attribute prompt for fine-grained I^{2}CL.

3.3 Fine-grained I^{2}CL

With photometric attribute classification results, we generate the fine-grained artifact spoof prompt $\boldsymbol{t}_D^+$ for I^{2}CL. For simplicity, if the attribute is classified as

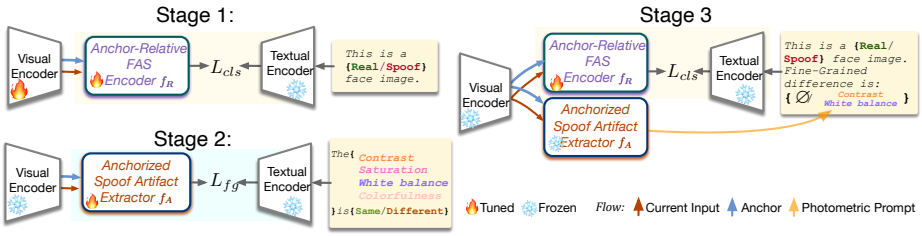

Fig. 3: The three-stage training strategy of the proposed TRAM.

'different', we concatenate ($\cup$) its corresponding textual prompt and the summation is:

$$\boldsymbol{t}_D^+ = \bigcup_{d \in D} (\boldsymbol{t}_d^+ \mid p(\hat{y}_d^+ | \boldsymbol{R}_A) > p(\hat{y}_d^- | \boldsymbol{R}_A)) \tag{9}$$

Eq. 2 of I^{2}CL is updated with $\boldsymbol{t}_s$ where $\boldsymbol{t}_s \leftarrow \boldsymbol{t}_s \cup \boldsymbol{t}_D^+$. For the spoof class, fine-grained $\boldsymbol{t}_s$ provides more deceptive information. It further boosts the discriminability of the f_R in detecting FAS-related photometric discrepancies through $\mathcal{L}_{cls}$. For the real class, we do not add supplement prompts. The final learning objective is organized as:

$$\mathcal{L}_{sum} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{car} + \lambda_2 \mathcal{L}_{fg} \tag{10}$$

where λ_1 and λ_2 balance the losses.

Inference. In the testing stage, TRAM pairs the inference input with a genuine face from the target domain as the anchor. FRPL first generates fine-grained spoofing artifact prompts according to the prediction of f_A and concatenates to the $\boldsymbol{t}_s$. The final decision is obtained with Eq. 2 by computing the similarity of relational feature with updated prompts. Different from popular DA methods that need model re-training with target domain real and attack samples, TRAM is finetuning-free since it is based on the relationship modeling of input image and real face anchor. In practice, we take a few shots of real faces from the target domain as anchor candidates to account for potential variations. An anchor is selected for each input using a simple intensity-based heuristic to roughly align the illumination condition between the pair. We empirically observe that TRAM is not sensitive to the specific anchor choice as long as the anchor is a genuine sample from the same domain.

3.4 Training strategy

There are two main learning objectives in the proposed TRAM: 1) learn discriminative FAS feature by modeling the relationship of the input image with real face anchor. 2) Mine low-level photometric discrepancy of the input pair to provide fine-grained spoof artifact prompts to describe the relation. Since the output of f_A is regarded as the text prompt guidance to finetune f_R, unreliable

Table 1: One-to-One cross-dataset evaluation between **M**, **I**, **C** and **O** (Protocol 1) with the assessment metrics HTER

Methods	Type	C → I	C → M	C → O	I → C	I → M	I → O	M → C	M → I	M → O	O → C	O → I	O → M	Avg.
ADDA [58]	DA	41.8	36.6	-	49.8	35.1	-	39.0	35.2	-	-	-	-	39.6
DRCN [13]	DA	44.4	27.4	-	48.9	42.0	-	28.9	36.8	-	-	-	-	38.1
DupGAN [16]	DA	42.1	33.4	-	46.5	36.2	-	27.1	35.4	-	-	-	-	36.8
KSA [25]	DA	39.3	15.1	-	12.3	33.3	-	9.1	34.9	-	-	-	-	24.0
DR-UDA [62]	DA	15.6	9.0	28.7	34.2	29.0	**3.5**	16.8	3.0	30.2	19.5	25.4	27.4	23.1
ADA [61]	DA	17.5	9.3	29.1	41.5	30.5	39.6	17.7	5.1	31.2	19.8	26.8	31.5	25.0
USDAN-Un [23]	DA	16.0	9.2	-	30.2	25.8	-	13.3	3.4	-	-	-	-	16.3
CDFTN-L [71]	DA	**1.7**	8.1	29.9	11.9	9.6	29.9	8.8	**1.3**	25.6	19.1	5.8	6.3	13.2
JDOT-FAS [42]	DA	3.3	5.4	-	24.4	5.8	-	4.4	**0.5**	-	-	-	-	7.3
GDA [82]	FTDA	15.10	5.8	-	29.7	20.8	-	12.2	2.5	-	-	-	-	14.4
FLIP-V [52]	VL-DG	15.08	13.73	12.34	4.30	9.68	7.87	0.56	3.96	4.79	2.09	5.01	6.00	7.12
FLIP-IT [52]	VL-DG	12.33	15.18	7.98	1.12	8.37	6.98	0.19	5.21	4.96	**0.16**	4.27	5.63	6.03
FLIP-MCL [52]	VL-DG	10.57	7.15	3.91	0.68	7.22	4.22	0.19	5.88	3.95	0.19	5.69	8.40	4.84
BUDoPT [33]	VL-DG	4.33	2.62	4.98	0.48	1.83	4.14	**0**	2.45	1.87	0.44	2.53	1.43	2.46
TF-FAS [65]	VL-DG	3.06	**1.59**	**3.78**	0.69	1.34	**2.50**	0.11	2.31	**1.40**	1.5	4.02	1.59	1.99
TRAM(ours)	VL-FTDA	3.05	1.90	4.44	**0.44**	**0.71**	3.33	**0**	1.30	1.68	0.44	**2.25**	**0.24**	**1.65**

DA: domain adaptation, TDA: Test-time DA, FTDA : finetuning free TDA, VL : Vision-Language-based.

f_A could cause the entire model to collapse. Therefore, we design a three-stage training strategy that updates f_R, f_A, and then f_R with fine-grained spoof artifact prompts sequentially.

Three Stage Training. As shown in Figure 3, in Stage 1 we tune the visual encoder and f_R with the standard FAS text prompt to let the model obtain basic FAS ability. In Stage 2, we freeze the visual encoder and f_R so that f_A could try its best to mine the low-level photometric discrepancy attribute from the visual embedding with the guidance of fine-grained photometric attribute pseudo label. In stage 3, given the predicted fine-grained spoof artifact prompts, we further finetune the f_R so that it can adapt to the variation of text guidance and achieve better generalizability. For each source domain, we randomly choose real face samples as anchors to form diverse anchor–input pairs.

4 Experiments

4.1 Experimental Setups

Databases, Protocols. We evaluate TRAM on nine public and widely used FAS databases: MSU-MFSD(**M**) [67], CASIA-FASD(**C**) [78], Idiap Replay attack(**I**) [9], OULU-NPU(**O**) [2], WMCA(**W**) [12], CASIA-CeFA(**C**) [27], CASIA-SURF(**S**) [76], HKBUMARsV2+(**H**) [34], and CelebA-Spoof((Cspf)) [77]. Five cross-dataset evaluation protocols are involved to cover diverse scenarios, with varying source domains and attack instruments. Following [52], we adopt the one-to-one and leave-one-out settings for **M**, **C**, **I**, and **O** under Protocol 1 (**P1**) and Protocol 2 (**P2**), and evaluate TRAM on large-scale FAS benchmarks, including **W**, **C**, and **S**, under Protocol 3 (**P3**). In Protocol 4 (**P4**), we evaluate TRAM on high-quality 3D mask attacks between **W** and **H**. In Protocol 5 (**P5**), we assess large-scale cross-dataset generalization in a single-source setting: the model is trained on CelebA-Spoof and evaluated separately on **M**, **C**, **I**, and **O**.

Table 2: Cross-dataset evaluation between **M**, **I**, **C** and **O** under Protocol 2.

Methods	Type	Venue	OCI→M			OMI→C			OCM→I			ICM→O			Avg. HTER
			HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	
SSAB-R [66]	DG	CVPR'22	6.67	98.75	-	10.00	96.67	-	8.88	96.79	-	13.72	93.65	-	9.80
LDCformer [19]	DG	ICIP'23	6.43	98.39	-	8.11	96.67	-	8.57	97.09	-	11.17	95.58	-	8.57
NAS-FAS [70]	DA	TPAMI'20	16.85	90.42	-	15.21	92.64	-	11.63	96.98	-	13.16	94.18	-	14.21
RFMeta [50]	DA	AAAI'20	13.89	93.98	-	20.27	88.16	-	17.30	90.48	-	16.45	91.16	-	16.97
D^2AM [7]	DA	AAAI'21	12.70	95.6	-	20.98	85.58	-	15.43	91.22	-	15.27	90.87	-	16.09
Self-DA [64]	TDA	AAAI'21	15.40	91.80	-	24.50	84.40	-	15.60	90.10	-	23.10	84.30	-	19.65
ANRL [30]	DA	ACMMM'21	10.83	96.75	-	17.85	89.26	-	16.03	91.04	-	15.67	91.90	-	15.09
ViTAF† [18]	DA	ECCV'22	2.92	99.62	91.66	1.40	99.92	98.57	1.64	99.64	91.53	5.39	98.67	76.05	2.83
SDA-FAS++ [38]	TDA	PAMI'24	3.75	98.92	-	1.11	99.82	-	**0.25**	**99.99**	-	3.23	99.65	-	2.09
WJDOT-FAS [42]	DA	IJCV'25	5.0	97.0	-	2.8	99.5	-	9.1	96.9	-	4.6	99.1	-	5.38
GDA [82]	FTDA	ECCV'22	9.20	98.00	-	12.20	93.00	-	10.00	96.00	-	14.40	92.60	-	11.45
TTDG-V [80]	FTDA	CVPR'24	4.16	98.48	-	7.59	98.18	-	9.62	98.18	-	10	96.15	-	7.84
OTA [26]	FTDA	CVPR'25	2.38	99.42	93.33	2.67	99.49	91.11	5.19	98.56	88.22	3.03	99.45	90.66	3.21
FLIP-MCL [52]	VL-DG	ICCV'23	4.95	98.11	74.67	0.54	99.98	**100.00**	4.25	99.07	84.62	2.31	99.63	92.28	3.01
CFPL-FAS [28]	VL-DG	CVPR'24	1.43	99.28	98.57	2.56	99.10	66.33	5.43	98.41	85.29	2.50	99.42	94.72	2.98
BUDoPT [33]	VL-DG	ECCV'24	0.40	99.99	99.67	0.26	99.96	99.92	1.38	99.69	**98.10**	1.60	99.51	97.18	0.91
TF-FAS [65]	VL-DG	ECCV'24	1.49	99.80	-	0.58	**99.99**	-	1.56	99.89	-	1.43	**99.93**	-	1.27
MVP-FAS† [69]	VL-DG	ICCV'25	1.44	99.85	98.53	0.33	99.93	98.64	4.02	99.00	71.21	5.17	98.16	61.69	2.74
GD-FAS [24]	VL-DG	ICCV'25	0	100	-	0.19	99.95	-	1.67	99.71	-	1.92	99.50	-	0.95
FLIP-MCL-5-shot [52]	VL-DA	ICCV'23	3.42	99.34	82.67	0.63	99.98	**100.00**	1.52	99.86	97.23	1.54	99.81	96.37	1.77
BUDoPT-5-Bshot [33]	VL-DA	ECCV'24	**0**	**100**	**100**	0.67	**99.99**	99.32	1.76	99.44	91.98	1.43	99.69	94.84	0.96
TRAM(ours)	VL-FTDA	ECCV'26	**0**	**100**	**100**	**0.11**	**99.99**	99.77	1.85	99.10	97.30	**1.27**	99.72	**98.35**	**0.81**

DA: domain adaptation, TDA: Test-time DA, FTDA : finetuning-free TDA, VL : Vision-Language-based, 5-Bshot: model incorporates five target-domain bonafide samples during training. †: Result obtained using the official implementation.

Baseline methods. We compare TRAM with 30 cross-domain FAS methods, including both popular and recently published SOTA approaches. Among them, 3 are finetuning-free test-time domain adaptation (FTDA) methods and 8 are CLIP-based vision–language (VL) methods.

Metrics. We follow [52] and evaluate the performance of our method with three metrics: Half Total Error Rate (HTER), Area Under the Receiver Operating Characteristic Curve (AUC) and True Positive Rate (TPR) at a False Positive Rate (FPR) 1% (TPR@FPR=1%).

Implementation Details. We use MTCNN [74] to extract the facial landmarks for alignment and crop it in size 224x224. For FAS photometric properties, we validate contrast, saturation, colorfulness and white balance inspired by [25]. We follow the first-stage training of [33] to pretrain f_v on the source dataset. f_R and f_A begin with RCA and ACA respectively, followed by two self-attention layers. The optimizer is Adam with a learning rate of 10^{-6}. We set $\{\alpha,\ \theta,\ \mu,\ \lambda_1,\lambda_2\}$ as $\{0.9,\ 0.9,\ 0.9,\ 1,\ 1\}$, respectively. For Consensus Attention Regularization, we update the $\bar{\boldsymbol{G}}_c$ after 100 iterations warm up and apply the $\mathcal{L}_{car}$ starting from 150 iterations. Following [52], we use CelebA-Spoof [77] as supplementary training data for **P1** and **P2**. For testing, we randomly sample five bonafide images from the target domain as anchor candidates. For each test sample, TRAM pairs it with one of the candidates based on image intensity similarity.

4.2 Evaluation

With limited variations in source domain (P1). Results in Table 1 show that TRAM can effectively learn relation modeling with limited source data and outperforms the SOTA method TF-FAS of 17%. For settings where **M** is the target dataset, TRAM achieves larger improvement (Avg:+37%). We also

Table 3: .Large-scale cross-dataset evaluation between **W**, **C** and **S** (Protocol 3)

Methods	CS→W			SW→C			CW→S			Avg.
	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER
ViT [18]	21.04	89.12	30.09	17.12	89.05	22.71	17.16	90.25	30.23	18.44
CFPL-FAS [28]	9.04	96.48	25.84	14.83	90.36	8.33	8.77	96.83	53.34	10.88
FGPL [17]	14.05	92.65	33.33	19.00	88.53	13.33	11.00	94.72	34.00	14.68
S-CPTL [14]	8.99	94.01	-	12.78	91.64	-	9.48	95.83	-	10.42
MVP-FAS† [69]	8.56	96.75	**82.19**	17.36	89.64	6.56	12.84	94.41	47.41	12.92
TRAM(ours)	**8.23**	**96.97**	55.33	**12.05**	**93.82**	**40.55**	**5.08**	**98.78**	**81.16**	**8.45**

†: Result obtained using the official implementation.

Table 4: 3D mask attack cross-dataset evaluation (Protocol 4)

Methods	W→H		
	HTER	AUC	TPR@ FPR=1%
FLIP-V [52]	2.43	99.61	91.67
FLIP-IT [52]	3.94	99.28	83.33
FLIP-MCL [52]	1.97	99.88	93.40
TRAM(ours)	**0.91**	**99.83**	**99.48**

Table 5: Large-Scale Single-Source cross-dataset evaluation (Protocol 5).

Methods	Cspf → **M**		Cspf → **C**		Cspf → **I**		Cspf → **O**		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER
ViTAF [18]	12.86	93.14	3.11	99.48	12.38	95.73	26.73	81.28	13.77
FLIP [52]	19.37	89.95	4.88	98.48	25.67	81.37	20.57	87.30	17.62
I-FAS [72]	5.63	**98.73**	1.11	99.88	9.15	98.73	14.86	92.68	7.69
TAR-FAS [73]	5.71	98.29	**0.00**	**100.00**	**5.75**	**98.29**	14.45	92.63	6.48
TRAM (ours)	**4.39**	98.39	0.56	99.89	9.58	96.02	**8.17**	**96.97**	**5.68**

note the slight performance drop for **C**→ others, as **C** contains salient spoofing cues (paper mask with obvious artifacts), which hinder the learning of more generalized FAS feature. CLIP-based methods outperform the other traditional methods in general, which validates the effectiveness of text-guidance in FAS although it tends to rely on low-level vision cues.

Leave-one-out cross-dataset evaluation (P2). In **P2**, TRAM outperforms the HTER of SOTA methods in all settings as shown in Table 2. Furthermore, our model performs well on the TPR@FPR=1% metric, which indicates its potential in real application. To ensure a fair comparison with identical target-domain information, BUDoPT-5-Bshot incorporates the same five target domain bonafide samples in training stage. While this yields gains in some settings, the improvement is not consistent. In contrast, TRAM achieves an average HTER of 0.81, indicating that its advantage comes from relation modeling rather than access to bonafide target samples alone. We can also observe the larger improvements compared with finetuning-free TTA methods: TTDG [80], GDA [82] and OTA [26], although TRAM uses only a few real faces from the testing domain.

Large-scale cross-dataset evaluation (P3). **P3** involves larger number of subjects, ethnic diversity and more diverse environmental variations, thereby providing a more realistic assessment of model performance. As shown in Table 3, TRAM consistently achieves the best overall performance, with the lowest average HTER and the highest AUC across all settings. TRAM also attains the largest gain on CW→S (+42%). This large-scale evaluation further validates TRAM's generalization and practical applicability.

High-quality 3D mask cross-dataset evaluation (P4). In **P4**, we use **W** with silicon 3D mask samples as the source domain and test on **H** with more

Table 6: Ablation study of major components in **P2** with metric HTER.

I^2CL	CAR	FRPL	OCI→M	OMI→C	OCM→I	ICM→O	Avg.
-	-	-	5.27	0.44	2.94	2.31	3.05
✓	-	-	0.24	0.67	4.05	1.87	1.70
✓	✓	-	0	0.33	3.55	1.41	1.32
✓	✓	✓	0	0.11	1.85	1.27	0.81

Table 7: Operator-level analysis of RCA in I^2CL with **P2** with metric HTER.

Variant	$\mathcal{T}(\cdot)$	OCI→M	OMI→C	OCM→I	ICM→O	Avg.
Siamese-based	No	0.71	1.44	8.75	3.74	3.66
Cross-attention baseline	No	1.45	0.79	5.90	3.05	2.80
RCA Q-difference only	No	0.97	0.66	4.66	3.32	2.40
RCA K-difference only	No	1.21	0.44	5.16	3.03	2.46
RCA $\boldsymbol{Q}/\boldsymbol{K}$	No	0.44	0.67	4.37	2.54	2.04
Full RCA	Yes	0.24	0.67	4.05	1.87	1.70

realistic 3D masks under diverse lighting conditions. As shown in Table 4, TRAM consistently outperforms the FLIP family baselines. We also observe from the failure cases that FLIP predictions are noticeably affected by lighting variations, whereas TRAM remains stable across different illumination conditions.

Large-Scale Single-Source cross-dataset evaluation (P5). We train TRAM on CelebA-Spoof and evaluate it on the heterogeneous OCIM domains, simulating deployment from a centralized source to unseen environments. As shown in Table 5, TRAM achieves the best average HTER among two MLLM methods I-FAS [72] and TAR-FAS [73], with the lowest HTER on **M** and **O** and competitive results on **C** and **I**. These results demonstrate effective transfer of liveness cues from a single source to unseen domains.

4.3 Ablation Study

Component-level analysis of TRAM. We evaluate the effectiveness of each component through the ablation study in Table 6 based on the settings of **P2**. FLIP-IT [52] is regarded as a baseline. Compared with FLIP, I^2CL introduces relationship modeling in a finetuning-free manner, reducing the overall HTER by 44%. With CAR, the model is encouraged to focus on more generalizable relational feature, which effectively improves its performance. The fine-grained prompts generated by FRPL enhance the model's sensitivity to low-level anomalies and reduce the average HTER by 38%. We report more qualitative analysis in the supplementary material to better illustrate the effect of component CAR and FRPL.

Operator-level analysis of I^2CL. Table 7 presents an operator-level analysis of RCA within I^2CL. We first implement an intuitive input–anchor interaction using a simple Siamese pair encoder, which independently encodes the two images, concatenates their features, and aligns the resulting representation with relation text embeddings. Replacing this naive fusion with a standard cross-attention encoder, in which input features query anchor features, reduces the average HTER from 3.66 to 2.80. RCA further incorporates anchor–input differences into both the query and key representations, achieving an average HTER of 2.04. Finally, mask-aware temperature modulation completes the full RCA design and further reduces the average HTER to 1.70.

Table 8: Effect of Anchor Selection Criteria in **P2**

Criteria	OCI→M	OMI→C	OCM→I	ICM→O	Avg. HTER
Semantic	0	0.44	2.00	1.42	0.97
Color	0	0.22	1.70	1.38	0.83
Intensity	0	0.11	1.85	1.27	0.81

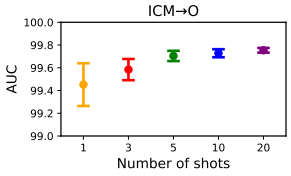

Fig. 4: Effect of #shots.

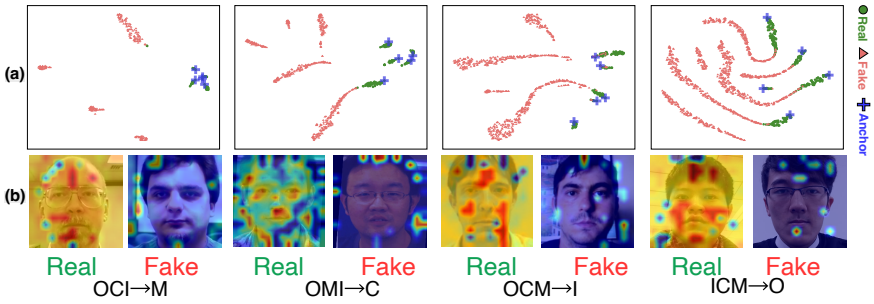

Fig. 5: (a)The t-SNE of TRAM relational features in the target domain under **P2**. (b) Typical attention maps on real and spoof images for different target domain in **P2**. Real faces show focused and coherent facial attention, while spoof faces exhibit scattered and inconsistent attention.

4.4 Hyperparameter Analysis

Effect of Different Anchor Selection Criteria. Table 8 evaluates the impact of different anchor selection strategy. Each criterion selects the closest anchor from the candidates. *Semantic* is based on feature similarity extracted by an ImageNet-pretrained ResNet-18. *Color* and *Intensity* rely on patch-wise RGB statistics and global intensity differences, respectively. *Intensity* achieves the lowest avg. HTER, suggesting that high-level semantic similarity is not necessary and may introduce task-irrelevant bias. The small performance differences across criteria demonstrate that TRAM is robust to the anchor selection strategy.

Effect of number of shots. Figure 4 reports the mean AUC and standard deviation on ICM→O under different numbers of real-face shots. For each shot setting, results are averaged over five independent trials with randomly sampled anchor candidates. Beyond 5 shots, the distribution changes only slightly, suggesting that TRAM becomes stable with a small number of shots. Five shots provide a good robustness efficiency trade off. Additional results on the effect of the number of shots under the remaining P2 settings are provided in the supplementary material.

4.5 Qualitative analysis

Visualization of Relational feature distribution. We plot the relational feature distributions of **P2** using t-SNE [41] in Figure 5 (a). Samples sharing the domain cluster together, and clear boundaries are shown between bonafide

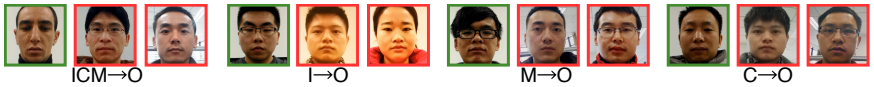

Fig. 6: Failure cases study of our method on target domain **O** with Protocol 1 and 2. Green boxes denote real samples and red boxes denote spoof samples.

and spoof samples. The anchors ("anchor-anchor" relational feature) denoted with blue cross tend to stay at the edge of the live clusters as their self-relations provide the highest confidence.

Attention Visualization. Figure 5 (b) presents typical Grad-CAM attention maps of the relational features on target-domain samples. TRAM exhibits a one-class classification–like behavior. For real faces, it focuses on facial regions and captures coherent relational feature patterns with the anchor. In contrast, spoof faces that differ from the anchor produce scattered and inconsistent attention.

Failure cases analysis. Figure 6 shows typical failures on target domain **O** with **P1** and **P2**. Green box denotes real samples and red denotes spoof ones. Most false rejections on real faces stem from low-quality inputs. An evident error pattern is observed in **I→O**, where warm-toned spoof samples are often misclassified as real. This can be attributed to a dataset bias in **I**, where warm-toned samples are all real. This issue is alleviated in **ICM→O**, indicating that larger source diversity reduces overfitting to salient shortcuts under limited variation.

5 Conclusion

In this work, we tackle the DA problem in FAS from a novel relative image-pair-text contrastive learning perspective. Since the real face image shares the same domain information as others, we take it as the anchor to model the relationship between the other inputs to conduct generalized FAS. The proposed TRAM achieves finetuning-free test-time DA with only a few target-domain genuine samples from it. A Relative Image-Pair-Text Contrastive Learning strategy is designed to learn discriminative spoof cues from the image pairs, guided by fine-grained FAS prompts and regularized by attention consensus. For lack of detailed spoof annotations, a Fine-grained Relation Prompt Learning strategy is proposed to mine FAS-related low-level photometric discrepancy information with pseudo detailed artifact attributes and generate fine-grained prompts. Extensive experiments on eight FAS datasets covering typical variations demonstrate the generalizability of TRAM over state-of-the-art methods. In-depth analysis of each TRAM component is also performed from multiple perspectives.

6 Acknowledgement

This work is supported by the Natural Science Foundation of Guangdong Province 2025A1515011556 and GuangDong Basic and Applied Basic Research Foundation 2023A1515110706. We thank Prof. Pong C. Yuen for insightful discussions and feedback that helped sharpen the positioning and presentation of this work.

References

1. Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: IASP (2009)
2. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: FG (2017)
3. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Trans. on Information Forensics and Security **11**(8), 1818–1830 (2016)
4. Cai, R., Cui, Y., Li, Z., Yu, Z., Li, H., Hu, Y., Kot, A.: Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. arXiv preprint arXiv:2303.09914 (2023)
5. Cai, R., Li, Z., Wan, R., Li, H., Hu, Y., Kot, A.C.: Learning meta pattern for face anti-spoofing. IEEE Transactions on Information Forensics and Security **17**, 1201–1213 (2022)
6. Chang, S., Ghamisi, P.: Changes to captions: An attentive network for remote sensing change captioning. IEEE Transactions on Image Processing (2023)
7. Chen, Z., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Huang, F., Jin, X.: Generalizable representation learning for mixture domain face anti-spoofing. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1132–1139 (2021)
8. Cheng, J., Liang, D., Tan, S.: Transfer clip for generalizable image denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25974–25984 (2024)
9. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: BIOSIG. pp. 1–7 (2012)
10. Dunlap, L., Zhang, Y., Wang, X., Zhong, R., Darrell, T., Steinhardt, J., Gonzalez, J.E., Yeung-Levy, S.: Describing differences in image sets with natural language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24199–24208 (2024)
11. Fang, H., Liu, A., Jiang, N., Lu, Q., Zhao, G., Wan, J.: Vl-fas: Domain generalization via vision-language model for face anti-spoofing. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4770–4774. IEEE (2024)
12. George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Trans. on Information Forensics and Security (2019)
13. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 597–613. Springer (2016)
14. Guo, J., Liu, H., Luo, Y., Hu, X., Zou, H., Zhang, Y., Liu, H., Zhao, B.: Style-conditional prompt token learning for generalizable face anti-spoofing. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 994–1003 (2024)
15. Guo, J., Zhu, X., Xiao, J., Lei, Z., Wan, G., Li, S.Z.: Improving face anti-spoofing by 3d virtual synthesis. arXiv (2019)
16. Hu, L., Kan, M., Shan, S., Chen, X.: Duplex generative adversarial network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1498–1507 (2018)

17. Hu, X., Liu, H., Yuan, H., Fu, Z., Luo, Y., Zhang, N., Zou, H., Gan, J., Zhang, Y.: Fine-grained prompt learning for face anti-spoofing. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7619–7628 (2024)
18. Huang, H.P., Sun, D., Liu, Y., Chu, W.S., Xiao, T., Yuan, J., Adam, H., Yang, M.H.: Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In: European Conference on Computer Vision. pp. 37–54. Springer (2022)
19. Huang, P.K., Chiang, C.H., Chong, J.X., Chen, T.H., Ni, H.Y., Hsu, C.T.: Ldc-former: Incorporating learnable descriptive convolution to vision transformer for face anti-spoofing. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 121–125. IEEE (2023)
20. Huang, P.K., Lu, C.Y., Chang, S.J., Chong, J.X., Hsu, C.T.: Test-time adaptation for robust face anti-spoofing. In: BMVC. pp. 379–380 (2023)
21. Jee, H.K., Jung, S.U., Yoo, J.H.: Liveness detection for embedded face recognition system. International Journal of Biological and Medical Sciences **1**(4), 235–238 (2006)
22. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8484–8493 (2020)
23. Jia, Y., Zhang, J., Shan, S., Chen, X.: Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. Pattern Recognition **115**, 107888 (2021)
24. Jung, S., Lee, K., Jeong, Y., Noh, H., Lee, J., Choi, J.: Group-wise scaling and orthogonal decomposition for domain-invariant feature extraction in face anti-spoofing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13372–13381 (2025)
25. Li, H., Li, W., Cao, H., Wang, S., Huang, F., Kot, A.C.: Unsupervised domain adaptation for face anti-spoofing. IEEE Transactions on Information Forensics and Security **13**(7), 1794–1809 (2018)
26. Li, Z., Zhao, T., Xu, X., Zhang, Z., Li, Z., Chen, X., Zhang, Q., Bergamo, A., Jain, A.K., Xing, Y.: Optimal transport-guided source-free adaptation for face anti-spoofing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24351–24363 (2025)
27. Liu, A., Tan, Z., Wan, J., Escalera, S., Guo, G., Li, S.Z.: Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1179–1187 (2021)
28. Liu, A., Xue, S., Gan, J., Wan, J., Liang, Y., Deng, J., Escalera, S., Lei, Z.: Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 222–232 (2024)
29. Liu, C., Zhao, R., Chen, J., Qi, Z., Zou, Z., Shi, Z.: A decoupling paradigm with prompt learning for remote sensing image change captioning. IEEE Transactions on Geoscience and Remote Sensing (2023)
30. Liu, S., Zhang, K.Y., Yao, T., Bi, M., Ding, S., Li, J., Huang, F., Ma, L.: Adaptive normalized representation learning for generalizable face anti-spoofing. In: Proceedings of the 29th ACM international conference on multimedia. pp. 1469–1477 (2021)
31. Liu, S., Zhang, K.Y., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Xie, Y., Ma, L.: Dual reweighting domain generalization for face presentation attack detection. arXiv preprint arXiv:2106.16128 (2021)

32. Liu, S.Q., Lan, X., Yuen, P.C.: Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. IEEE Trans. on Information Forensics and Security **16**, 2683–2696 (2021)
33. Liu, S.Q., Wang, Q., Yuen, P.C.: Bottom-up domain prompt tuning for generalized face anti-spoofing (2024)
34. Liu, S., Yang, B., Yuen, P.C., Zhao, G.: A 3d mask face anti-spoofing database with real world variations. In: CVPRW (2016)
35. Liu, Y., Liu, X.: Spoof trace disentanglement for generic face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3813–3830 (2022)
36. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: CVPR (2019)
37. Liu, Y., Stehouwer, J., Liu, X.: On disentangling spoof trace for generic face anti-spoofing. In: European Conference on Computer Vision. pp. 406–422. Springer (2020)
38. Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C.T., Xiong, H.: Source-free domain adaptation with domain generalized pretraining for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5430–5448 (2024)
39. Lv, L., Xiang, Y., Li, X., Huang, H., Ruan, R., Xu, X., Fu, Y.: Combining dynamic image and prediction ensemble for cross-domain face anti-spoofing. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2550–2554. IEEE (2021)
40. Ma, Y., Lin, X., Xu, Y., Xie, W., Yu, Z.: Pa-fas: Towards interpretable and generalizable multimodal face anti-spoofing via path-augmented reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
41. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
42. Mao, S., Chen, R., Li, H.: Weighted joint distribution optimal transport based domain adaptation for cross-scenario face anti-spoofing. International Journal of Computer Vision **133**(2), 590–610 (2025)
43. Pan, G., Sun, L., Wu, Z., Lao, S.: Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: ICCV. pp. 1–8 (2007)
44. Patel, K., Han, H., Jain, A.K.: Secure face unlock: Spoof detection on smartphones. IEEE transactions on information forensics and security **11**(10), 2268–2283 (2016)
45. Qin, Y., Zhao, C., Zhu, X., Wang, Z., Yu, Z., Fu, T., Zhou, F., Shi, J., Lei, Z.: Learning meta model for zero-and few-shot face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11916–11923 (2020)
46. Qiu, Y., Yamamoto, S., Nakashima, K., Suzuki, R., Iwata, K., Kataoka, H., Satoh, Y.: Describing and localizing multiple changes with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1971–1980 (2021)
47. Quan, R., Wu, Y., Yu, X., Yang, Y.: Progressive transfer learning for face anti-spoofing. IEEE Transactions on Image Processing **30**, 3946–3955 (2021)
48. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
49. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10023–10031 (2019)

50. Shao, R., Lan, X., Yuen, P.C.: Regularized fine-grained meta face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11974–11981 (2020)
51. Shi, Y., Gao, Y., Lai, Y., Wang, H., Feng, J., He, L., Wan, J., Chen, C., Yu, Z., Cao, X.: Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. arXiv preprint arXiv:2402.04178 (2024)
52. Srivatsan, K., Naseer, M., Nandakumar, K.: Flip: Cross-domain face anti-spoofing with language guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19685–19696 (2023)
53. Sun, X., Hu, P., Saenko, K.: Dualcoop: Fast adaptation to multi-label recognition with limited annotations. Advances in Neural Information Processing Systems **35**, 30569–30582 (2022)
54. Sun, Y., Liu, Y., Liu, X., Li, Y., Chu, W.S.: Rethinking domain generalization for face anti-spoofing: Separability and alignment. In: CVPR. pp. 24563–24574 (June 2023)
55. Tan, X., Li, Y., Liu, J., Jiang, L.: Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: ECCV (2010)
56. Tu, Y., Li, L., Su, L., Lu, K., Huang, Q.: Neighborhood contrastive transformer for change captioning. IEEE Transactions on Multimedia **25**, 9518–9529 (2023)
57. Tu, Y., Li, L., Su, L., Zha, Z.J., Yan, C., Huang, Q.: Self-supervised cross-view representation reconstruction for change captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2805–2815 (2023)
58. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017)
59. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
60. Wang, C.Y., Lu, Y.D., Yang, S.T., Lai, S.H.: Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20281–20290 (2022)
61. Wang, G., Han, H., Shan, S., Chen, X.: Improving cross-database face presentation attack detection via adversarial domain adaptation. In: 2019 International Conference on Biometrics (ICB). pp. 1–8. IEEE (2019)
62. Wang, G., Han, H., Shan, S., Chen, X.: Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. IEEE Transactions on Information Forensics and Security **16**, 56–69 (2020)
63. Wang, H., Shi, Y., Tao, Z., Gao, Y., Zhang, L., Lin, X., Feng, J., Yuan, X., Yu, Z., Cao, X.: Faceshield: Explainable face anti-spoofing with multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
64. Wang, J., Zhang, J., Bian, Y., Cai, Y., Wang, C., Pu, S.: Self-domain adaptation for face anti-spoofing. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 2746–2754 (2021)
65. Wang, X., Zhang, K.Y., Yao, T., Zhou, Q., Ding, S., Dai, P., Ji, R.: Tf-fas: twofold-element fine-grained semantic guidance for generalizable face anti-spoofing. In: European Conference on Computer Vision. pp. 148–168. Springer (2025)
66. Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., Wang, Z.: Domain generalization via shuffled style assembly for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4123–4133 (2022)
67. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Trans. on Information Forensics and Security **10**(4), 746–761 (2015)

68. Yang, H., Pan, L., Yang, Y., Hartley, R., Liu, M.: Ldp: Language-driven dual-pixel image defocus deblurring network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24078–24087 (2024)
69. Yu, J., Kim, S., Lee, K., Kwon, T., Shin, W.Y., Kim, H.Y.: Multi-view slot attention using paraphrased texts for face anti-spoofing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21117–21128 (2025)
70. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: Nas-fas: Static-dynamic central difference network search for face anti-spoofing. IEEE Trans. on Pattern Analysis and Machine Intelligence (2020)
71. Yue, H., Wang, K., Zhang, G., Feng, H., Han, J., Ding, E., Wang, J.: Cyclically disentangled feature translation for face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3358–3366 (2023)
72. Zhang, G., Wang, K., Yue, H., Liu, A., Zhang, G., Yao, K., Ding, E., Wang, J.: Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9896–9904 (2025)
73. Zhang, H., Wang, K., Zhang, G., Yue, H., Tan, Z., Peng, S., Zhang, T., Tan, X., Chen, K., He, W., et al.: From intuition to investigation: A tool-augmented reasoning mllm framework for generalizable face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 40855–40865 (2026)
74. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters **23**(10), 1499–1503 (2016)
75. Zhang, R., Zeng, Z., Guo, Z., Li, Y.: Can language understand depth? In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6868–6874 (2022)
76. Zhang, S., Liu, A., Wan, J., Liang, Y., Guo, G., Escalera, S., Escalante, H.J., Li, S.Z.: Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science **2**(2), 182–193 (2020)
77. Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., Liu, Z.: Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: European Conference on Computer Vision (ECCV) (2020)
78. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: ICB (2012)
79. Zhao, C., Wang, K., Hsiao, J.H., Chan, A.B.: Grad-eclip: Gradient-based visual and textual explanations for clip. arXiv preprint arXiv:2502.18816 (2025)
80. Zhou, Q., Zhang, K.Y., Yao, T., Lu, X., Ding, S., Ma, L.: Test-time domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 175–187 (2024)
81. Zhou, Q., Zhang, K.Y., Yao, T., Lu, X., Yi, R., Ding, S., Ma, L.: Instance-aware domain generalization for face anti-spoofing. In: CVPR. pp. 20453–20463 (June 2023)
82. Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Sheng, K., Ding, S., Ma, L.: Generative domain adaptation for face anti-spoofing. In: European Conference on Computer Vision. pp. 335–356. Springer (2022)
83. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)